# El Proyecto SCASEM
# un Sistema de Catalogación Semántica

**David Trotzig**
Planeta Actimedia, S.A.
Banco de Contenidos
Departamento de Lingüística Computacional
mailto:dtrotzig@planeta-actimedia.es

**Resumen** Para dar respuesta a las nuevas demandas del mercado de la publicación electrónica, el Grupo Planeta ha iniciado un proyecto con el propósito de adaptar la gestión de su fondos enciclopédicos, iconográficos, cartográficos y audiovi-suales al nuevo mercado, el Banco de Contenidos Planeta. Como núcleo clasifi-cador de este repositorio general se ha diseñado un sistema basado en las más avanzadas tecnologías tanto a nivel de redes semánticas, como de creación y aplicación de herramientas vinculadas al procesamiento del lenguaje natural (PLN).

## *1. El Banco de Contenidos Planeta*

En Planeta Actimedia se gestiona el Banco de Contenidos Planeta (BCP). El Banco de Contenidos es la base para la creación de productos multicanal y se organiza en fondos enciclopédicos, iconográficos, cartográficos y audiovisuales. El BCP está organizado y codificado mediante el uso de sistemas de gestión de contenidos basados en SGML y XML, sistemas de catalogación automática y de búsqueda por lenguaje natural.
Los principales objetivos del BCP son los siguientes:

- La clasificación de forma unificada de todo el material documental:
  Contenidos textuales: definiciones, artículos enciclopédicos, etc.
  Contenidos no textuales: fotografías, tablas, taxonomías, cartografía, etc.
  Contenidos audiovisuales:, ilustraciones, animaciones, vídeos, etc.
  Relaciones: árboles temáticos, árboles históricos, etc.
- La reutilización de material tanto documental como de conocimientos.
- Evitar que material se pierda o se vuelva inutilizable. Aquí algunos de los objetivos son:
  Evitar la desclasificación de material: poder encontrar cualquiera de los elementos de BCP con la misma facilidad.
  Evitar que el material quede desfasado: poder poner la día todo material que lo necesite.
- Evitar la necesidad de reclasificar material a medida que vaya creciendo el Banco de Contenidos:
  Siendo el repositorio general de una gran cantidad de material textual o de otro tipo, necesita un proceso de realimentación tanto del material nuevo como del material que ha sido modificado o mejorado para una obra determinada.

El BCP consta de tres partes generales:

1) Los fondos enciclopédicos, iconográficos, cartográficos y audiovisuales del banco.
2) La Base de Datos de Conocimiento (BDCon), que contiene una ontología multidimensional y que constituye la herramienta central del sistema de clasifi-cación del banco.
3) Un conjunto de herramientas lingüísticas dedicadas al tratamiento automático o semiautomático de los contenidos textua-les del Banco de Contenidos.

## *2. El proyecto SCASEM*

El propósito del proyecto **SCASEM** (Sistema de Catalogación Semántica), es la creación de la base de conocimiento (la BDCon), que constituye el núcleo del sistema, y el desarrollo del conjunto de herramientas lingüísticas que darán apoyo a los editores de las obras del futuro.

### *2.1. La BDCon*

La BDCon tiene la función de clasificar los contenidos del BCP de forma que sus usuarios puedan encontrar todo el material que buscan, sea textual, fotográfico u otro, con un máximo de precisión y exactitud. También sirve de apoyo semántico para las herramientas de tratamiento de texto.

### *2.2. La BDCon como clasificador/buscador*

La BDCon es una ontología general que está organizada alrededor de una columna vertebral compuesta por un **lemario** muy extenso (más de 1.000.000 entradas) conectado con un diccionario de raíces, y una serie de estructuras de conocimiento, cada una cubriendo un aspecto específico del material de base. Se puede ver como un sistema de coordenadas donde cada elemento del Banco, trátese de unidades lexicales, nombres propios o de documentos de cualquier tipo, puede ser identificado de forma unívoca. Los parámetros utilizados fueron:

1) El tipo de entidad del que se trata, organizado el un árbol tipológico (**ATIP**).
2) El área temática a la que pertenece la entidad, organizado en un árbol temático (**ATEM**).
3) El tipo de entidad desde el punto de vista del continente, no del contenido (tipo de palabra o tipo de documento) tomando en cuenta el soporte en que se encuentra. Organizado en un árbol de tipos de soporte (**ASOP**).
4) El lugar al que se puede asociar la entidad, organizado en una estructura geográfica (**EGEO**).
5) El período temporal o fecha al que se puede asociar la entidad, organizado en una estructura cronológica (**ECRON**).

Además, para tener un lugar donde relacionar las entidades entre sí, se ha creado una estructura llamada **RelCon**, de Relaciones de Conocimiento donde se relacionan, en principio, objetos concretos como personas, obras, lugares, conceptos, etc.

## *3. Las herramientas del Banco de Contenidos Planeta*

### 3.1. Función de las herramientas

- La recuperación “inteligente” de información.
- La catalogación automática de documentos.
- La creación automática de hipertexto.
- La comprobación y corrección automática de textos.
- El control automático de contenidos desde el punto de vista de las actualizaciones.
- La codificación automática de textos..
- La creación automática de resúmenes.
- La extracción de información y conocimiento de los contenidos del BCP.
- La creación de un sistema de búsqueda en lenguaje natural para el usuario final.

### 3.2. Las herramientas por módulos

El Lematizador: un tagger con una desambiguación basada los modelos Markov.
Chap: un parser tabulador que trabaja con una gramática general del castellano.
El Lector Semántico: un desambiguador semántico que se basa en inferencias sobre las relaciones entre nodos de los árboles de la BDCon para asignar un índice de probabilidad entre ambigüedades semánticas.
El Corrector de Planeta: un corrector orto tipográfico, gramatical y de estilo que ha sido desarrollado principalmente para el uso interno de las empresas del grupo Planeta.
El marcador de hipertexto: una aplicación que usa el *Lector Semántico* para crear marcas de hipertexto ponderadas según el contexto de la frase.
El Tematizador: un sistema de clasificación semántica de fotografías a partir de su pie de foto.
El Resumidor: un sistema de creación de resúmenes de artículos enciclopédicos.